Resource Management Corpus:
September 1992 Test Set Benchmark Test Results

David S. Pallett, Jonathan G. Fiscus and John S. Garofolo
National Institute of Standards and Technology (NIST)
Building 225 (Technology) Room A216
Gaithersburg, MD 20899

Abstract

This paper documents a set of Resource Management Corpus Benchmark Tests conducted prior to the DARPA ANN Technology Program Continuous Speech Recognition Meeting at Stanford University, September 21-22, 1993. Three organizations participated in tests of Speaker Dependent systems, reporting word errors using the word-pair grammar ranging from 1.8% to 2.6%. Eleven research teams from nine organizations reported results for a total of 18 Speaker Independent systems, with word error rates ranging from 4.4% to 11.7%. Properties of the September 1992 test sets are discussed, particularly with regard to the distribution of speaking rate for the speaker independent test set population. Two DARPA ANN Technology Program contractors reported results for both baseline HMM and hybrid ANN-HMM systems. The use of newly-implemented statistical significance tests underscores the need to interpret these test results on a case-by-case basis.

1. Introduction

The DARPA Resource Management Continuous Speech Database (RM1) [1,2] has been used for development and performance evaluation of automatic speech recognition technology within the DARPA speech research community since 1987.

DARPA-sponsored benchmark tests of both speaker-dependent and speaker-independent technology have been conducted in March and October 1987, June 1988, February and October 1989, February 1991, and most recently, September 1992. Additional benchmark tests were conducted in June, 1990, using the "Extended Resource Management" corpus (RM2), which contains extensive additional training material for (only) 4 speakers, to permit training of speaker-dependent systems with significantly more (2400 vs. 600) utterances than is possible with the RM1 corpus. While the tests have been "locally implemented" by system developers, NIST has provided the sponsors and the community the service of uniformly scoring results and providing summaries of the results, usually in conjunction with DARPA speech research meetings. For some of the tests, only informal "handouts" were prepared for distribution at DARPA workshops, but the March 1987, February 1989, June 1990, and February 1991 tests are documented [3-7].

This paper documents the most recent tests, conducted prior to the September 1992 DARPA Artificial Neural Network (ANN) Technology Program Meeting, using the final, previously unreleased, RM1 test set. Since, in the 5 years since release of the first test set, many researchers have developed "resource management" continuous speech recognition systems, the opportunity to participate in these tests was extended to a number researchers

outside of the DARPA community. Two DARPA ANN Technology Program contractors (BBN and SRI International) participated, along with a number of other research teams, including three from Europe (Philips Research Laboratories, in Aachen, Germany, CNRS-LIMSI, in Paris, France, and Cambridge University, in Cambridge, UK).

2. Test Protocols

The September 1992 test protocols followed well established precedents.

All sites had had access to all previously released training data, test sets, and "official" scoring software for some time. The September 1992 test sets were distributed on CD-ROM media on August 24th, results were reported to NIST on September 4th, and all scored results (including statistical significance tests) were made available to the participants via ftp on September 16th.

During the period between September 4th and 16th, some sites reported that the test data seemed "different" from prior test sets (i.e., unusually challenging), and NIST initiated analysis of the test sets to ascertain possible sources of the differences. The results of these analyses are indicated in the following section.

3. Properties of the September 1992 Test Set

Are these test sets different?

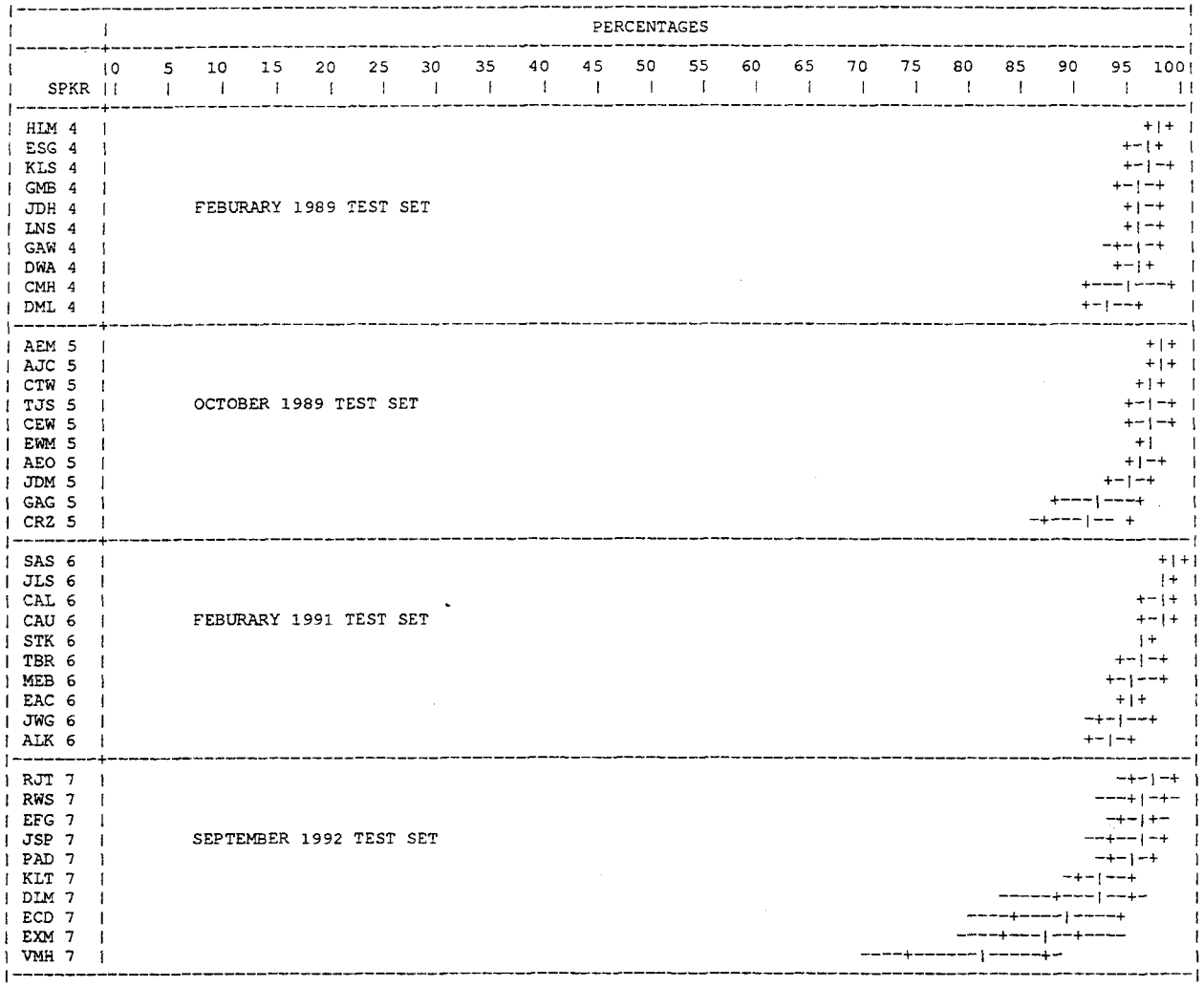
Researchers at BBN, Cambridge University, and CNRS-LIMSI provided results for the results of four different systems having processed the February and October 1989, February 1991, and the September 1992 Speaker Independent test sets, in order to illustrate differences in performance for several systems on four different test sets, involving a total of 40 speakers.

Figure 1 shows the results for these systems and test sets. For each subject in each test set, the range, mean, and the interval bounded by one standard deviation around the mean word error rate for the four systems are shown. In each test subset, results are ordered from best subject (the so-called "sheep") to worst ("goat"). In many cases, for "goats" poor performance is correlated with large variability across the several systems.

For example, for subject HLM in the February 1989 test set, the mean word accuracy was approximately 98% with a small standard deviation, while for subject CMH, the mean word error was approximately 95% with a significantly larger standard deviation.

Contrasting results for the current test set with the others, it can be seen that there are more "goats" in this test set than in previous test sets. Note, for example, that for subject VMH, the speaker with worst performance, the range of data extends from 67% to about 87%. Six of the ten speakers in this test set appear to have unusual performance, in the sense of having higher error rates than their "peers" and larger across-system ranges. These speakers are PAD, KLT, DLM, ECD, EXM, and VMH.

SPEAKER WORD ACCURACY RANGE ANALYSIS ACROSS SPEAKERS FOR THE
Feb 89, Oct 89, Feb 91 and Sep 92 Test Sets
Grouped by Test Set



| -> shows the mean
+ -> shows plus or minus one standard deviation

Figure 1

Interpretation of results for earlier test sets may be somewhat compromised by the fact that developers might have "tuned" their systems to the properties of earlier tests, but most of the sites indicated that that was not the case.

How are these test sets different?

Since there is some evidence that the current test set population has an unusual number of speakers for whom their speech is unusually difficult to recognize, it is of interest to speculate how that might have occurred, and what are the factors contributing to the unusual degree of difficulty.

A partial answer may lie in the history of the test set selection process.

At the outset of this benchmark test process (ca. 1986-1987), the test sets were partitioned by NIST from larger test sets. For speaker independent technology, 4 sets of 10 speakers were defined from the original 40 speaker evaluation test set population described in [1], and a conscious effort was made to maintain a consistent balance of dialect region and gender for these 4 sets. In retrospect, this effort was at least partially flawed in that some sets were better-balanced than others.

The most recent test set -- defined, in some sense, as the residue of having reasonably well-balanced earlier test sets -- has a larger fractional representation of female speakers than most earlier sets. It also has no speakers from the North Midland or South Midlands dialect region, with disproportionate representation of other dialect regions. For the speaker dependent test set, efforts were made to overcome possible "within session effects", but it is possible that some systematic effects were still encountered.

Most of NIST's analyses (including such properties as the number of words in each test set, the number of lexemes, the average number of words/utterance, the test set perplexities, the mean speech-to-noise ratio, and mean duration and speaking rate) do not seem to indicate any particularly unusual properties for September 1992 test set. These properties are summarized in Table 1, showing lexical properties of the several RM1 test sets, and Table 2, showing speech signal and speaker population properties.

Analyses of performance results for all test sets by dialect region and gender do not seem particularly informative in the sense that no one dialect region or gender seemed particularly difficult. This would tend to discredit the hypothesis that the source of the unusual difficulty was related to disproportionate representation of females and the absence of test speakers with North Midland or South Midland dialect.

It was suggested that there may be a correlation between poor performance and rate of speech. Estimates of rate of speech are possible by counting the number of words uttered and determining the utterance duration. For the RM1 test material this ranges from approximately 100 to 200 words/minute.

Lexical Properties of the RM1 Evaluation Test Sets

	Feb. '89 SI	Feb. '89 SD	Oct. '89 SI	Oct. '89 SD	Feb. '91 SI	Feb. '91 SD	Sep. '92 SI	Sep. '92 SD
Words	2561	2522	2684	2608	2484	2600	2559	2558
Lexemes	577	446	577	438	547	445	609	433
Sentences	300	300	300	300	300	300	300	300
Unique Prompts	270	150	240	150	240	150	270	150
Avg. Words/Utt.	8.5	8.4	9.0	8.7	8.3	8.7	8.5	8.5
Perplexity	62.26	61.35	58.32	59.21	61.33	61.00	61.50	61.71

Table 1

Speech Signal and Speaker Population Properties of the RM1 Evaluation Test Sets

	Feb. '89 SI	Feb. '89 SD	Oct. '89 SI	Oct. '89 SD	Feb. '91 SI	Feb. '91 SD	Sep. '92 SI	Sep. '92 SD
Mean Speech to Noise (dB.)	48.81	45.36	48.54	45.10	48.66	45.19	49.61	45.60
Mean Duration (sec.)	3.31	3.25	3.43	3.28	3.29	3.19	3.29	3.16
Percent Speech	96.93	97.65	95.81	97.48	96.50	98.34	97.12	97.77
Mean Speaking Rate (w.p.m.)	154.7	155.2	156.5	159.0	151.0	163.0	155.6	161.8
Speaker Gender	6M/4F	7M/5F	7M/3F	*	5M/5F	*	5M/5F	*
Speaker Dialect	NE: 2 N: 1 NM: 1 SM: 1 S: 2 NY: 1 W: 1 AB: 1	NE: 1 N: 2 NM: 2 SM: 2 S: 2 NY: 1 W: 2 AB: 0	NE: 1 N: 2 NM: 1 SM: 1 S: 1 NY: 1 W: 2 AB: 1	*	NE: 2 N: 1 NM: 2 SM: 1 S: 3 NY: 0 W: 1 AB: 0	*	NE: 3 N: 2 NM: 0 SM: 0 S: 3 NY: 1 W: 1 AB: 0	*

* same speaker set as Feb. '89 SD

Figure 2 shows plots, for the four systems, of measured word error against speaking rate for the 40 speakers in the 4 test sets. Note that the word error rate is high for the two fastest speakers, VMH and KLT, and that there is also large variability across systems for DLM, another faster-than-average speaker. A speaker with unusually slow speaking rate, EXM, has high word error rate and high variability across systems. So there is some evidence to support the hypothesis that rate of speech may be a factor contributing to poor performance, and that the September 1992 speaker-independent test set has an unusually high number of other-than-average-speaking rate speakers.

Note that some systems have more difficulty in dealing with these faster- and slower-than-average speakers than other systems.

It has been suggested that one contributing factor may be inadequate representation of varied rate of speech in the training material. Figure 3 shows histograms indicating the distribution of speaking rate for (a) the SI-109 training set, with 109 speakers, and (b) the 40 speakers in the four recent test sets. Note that both distributions have similar means (approximately 155 words per minute) and are broad (standard deviations of 20 and 15.7 words per minute), with some representation of both fast and slow speakers in both sets. Thus system developers have had access to training material for fast and slow speakers. However we are not as much concerned with the entire set of 40 speakers in the 4 test sets as we are concerned with the differences between the 4 test sets.

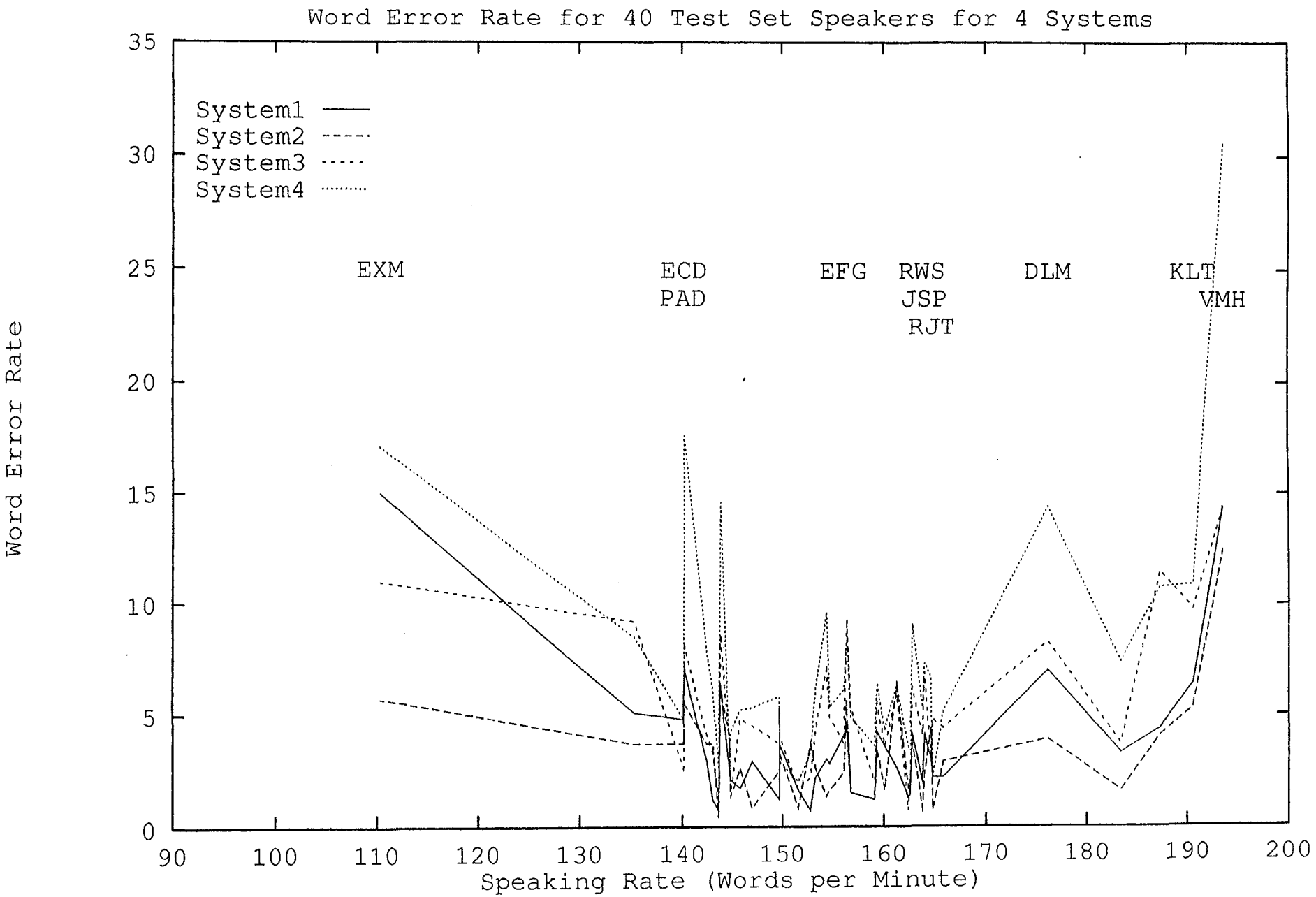
Figure 4 shows histograms for the speaking rate for the speakers in each of the four test sets. Note that while the mean speaking rate is similar for all four test sets (roughly 150 - 160 words per minute), the September 1992 test set is unusual in that it contains more "outliers" than previous test sets, with a larger standard deviation than other test sets. Note also, that the most recent previous test set had a smaller than usual standard deviation.

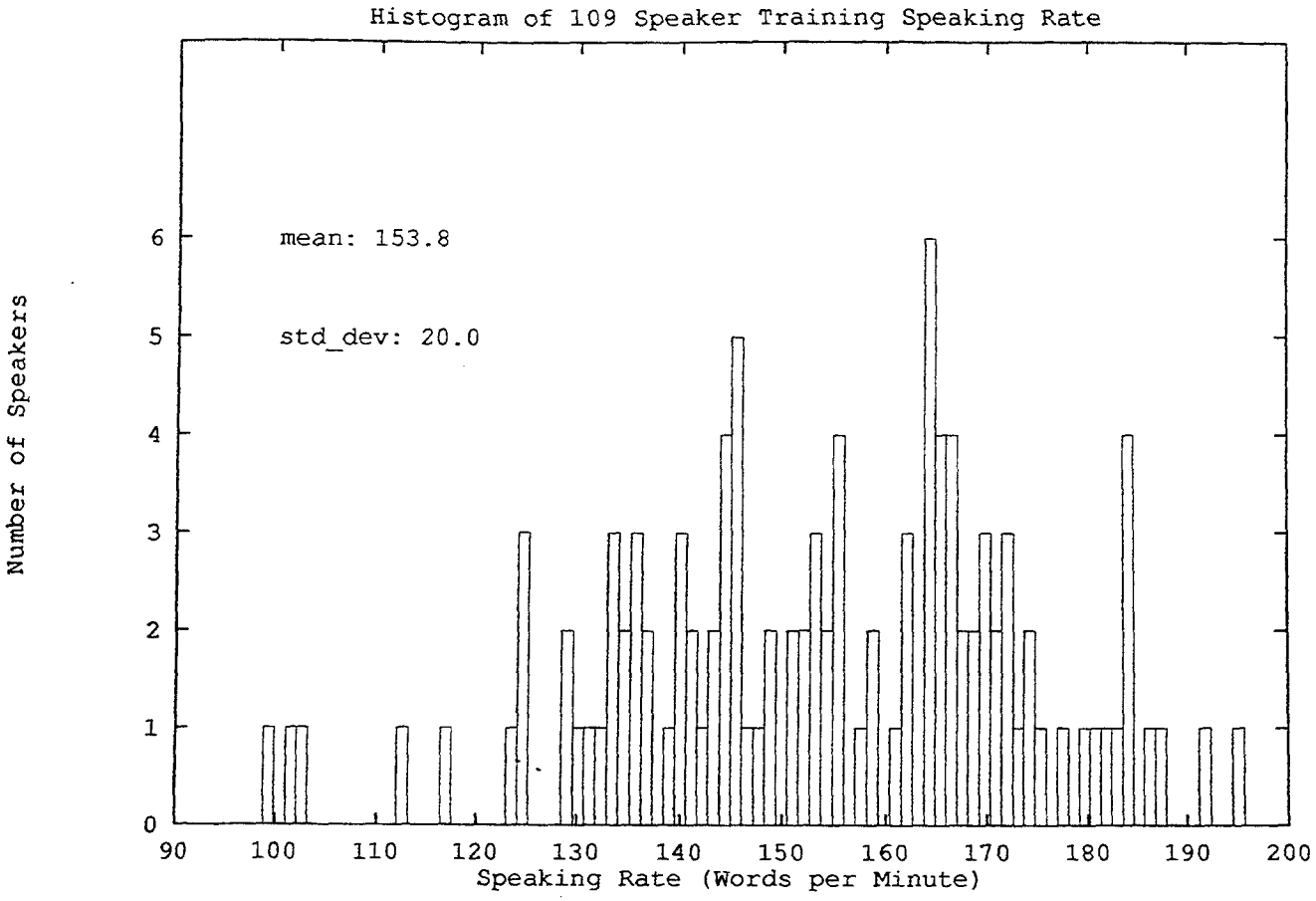
NIST's procedures for selection of test material have not involved "screening" potential test material to identify "outliers" of this sort (i.e., on rate of speech), but has relied principally on considerations of gender and dialect-region. Thus these differences in the properties of the test sets are entirely inadvertant. A case may be made for introducing screening proceedures in selection of future test sets, but it is difficult to identify all potential factors that might contribute to differences in "degree of difficulty".

Does it matter that this test material is unusually difficult?

There are a number of valid answers to this question, positive and negative, depending on one's understanding of the purpose(s) of the test.

Yes. The fact that this test is more difficult than others certainly complicates interpretation of "trends" in development of this technology, since the error rates for these test sets are in most cases higher than the best error rates cited on previous sets. Ideally, all test sets should be of equal difficulty, and technology improvements would be indicated directly by reductions in error rates.


Word Error Rate for 40 Test Set Speakers for 4 Systems
System1
System2
System3
System4
EXM
ECD
PAD
EFG
RWS
JSP
RJT
DLM
KLT
VMH
Word Error Rate
Speaking Rate (Words per Minute)
0
5
10
15
20
25
30
35
90
100
110
120
130
140
150
160
170
180
190
200



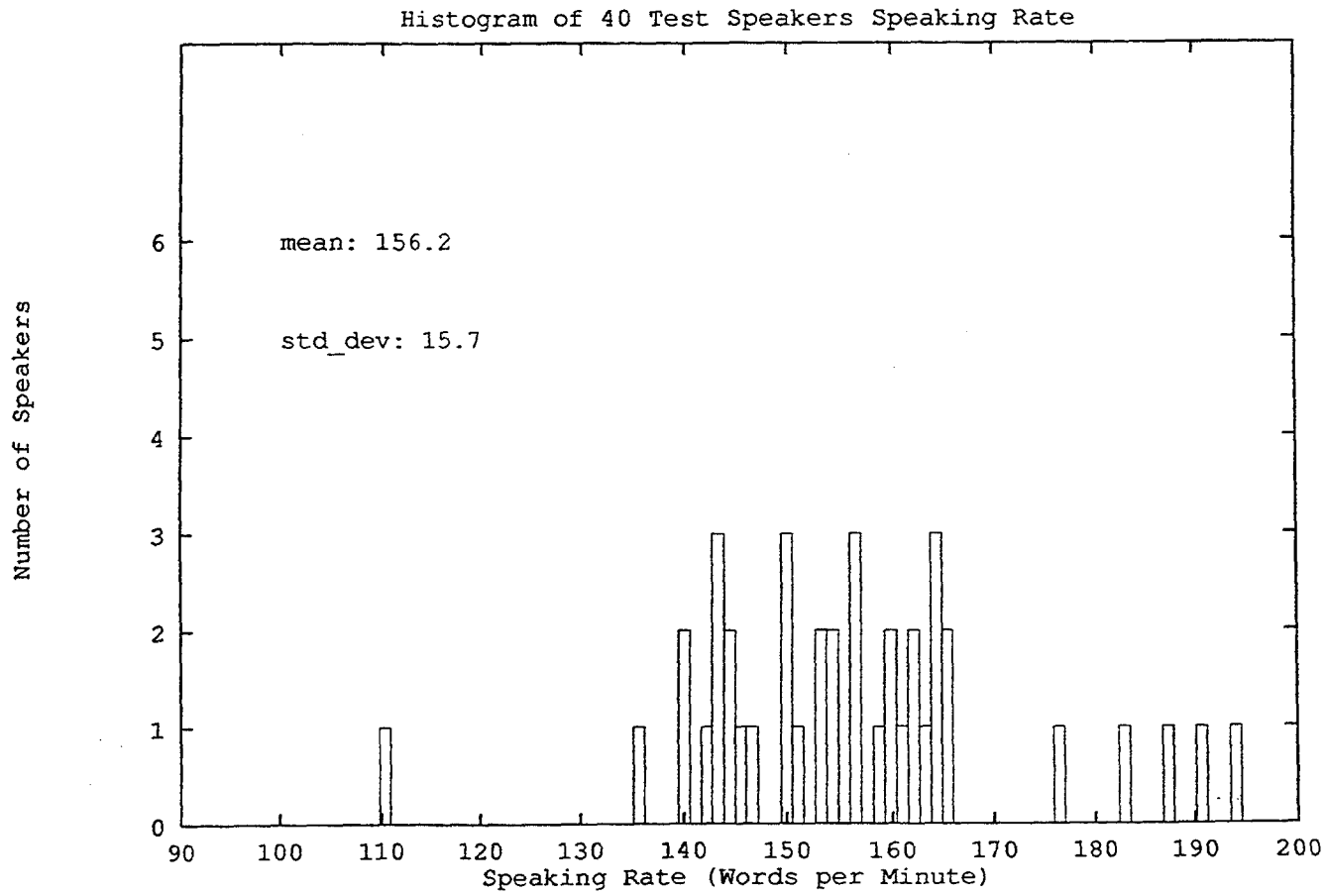


Figure 3: A (top) and B (bottom)

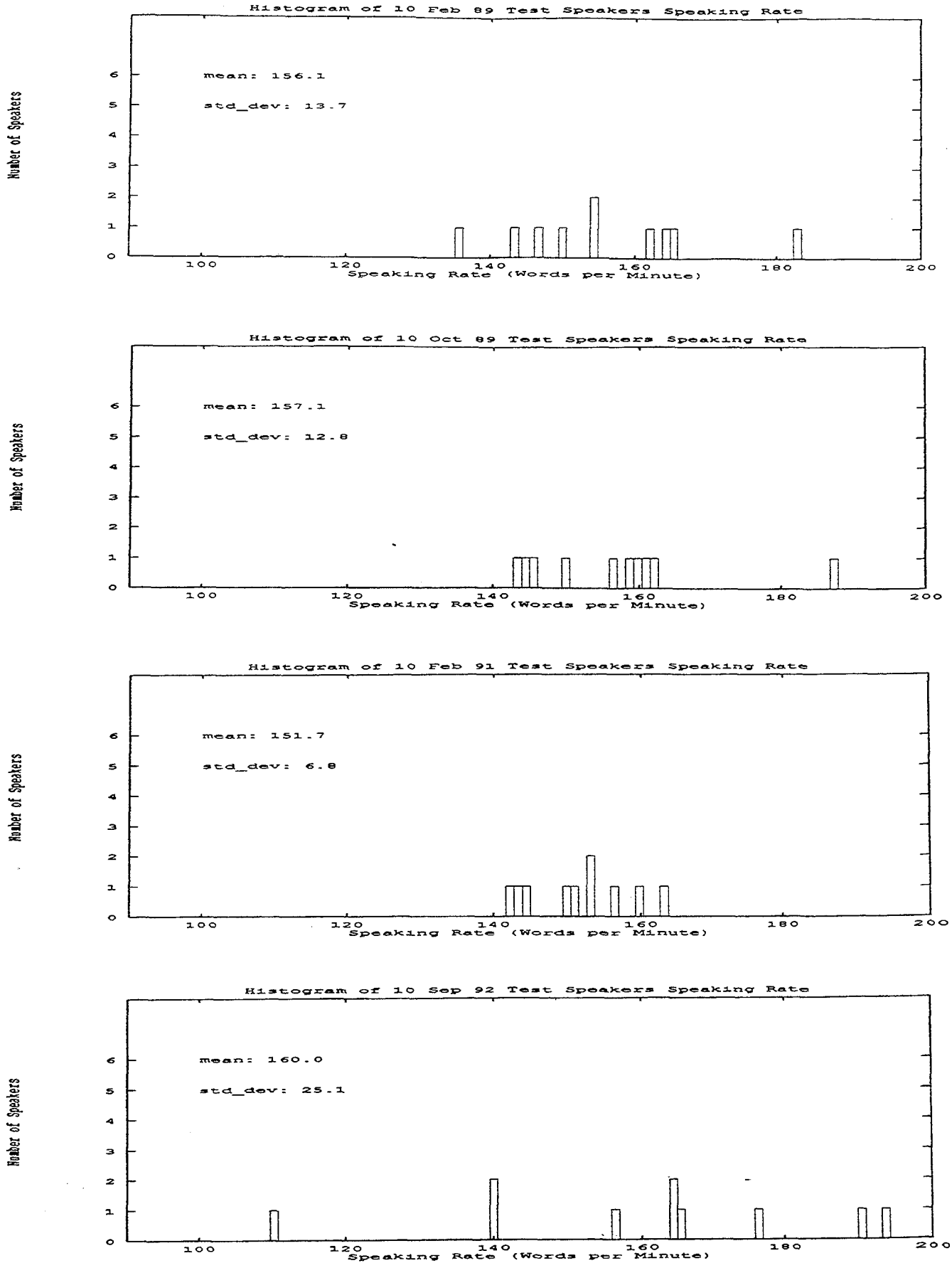

Histogram of 10 Feb 89 Test Speakers Speaking Rate
mean: 156.1
std_dev: 13.7
Number of Speakers
Speaking Rate (Words per Minute)
Histogram of 10 Oct 89 Test Speakers Speaking Rate
mean: 157.1
std_dev: 12.8
Number of Speakers
Speaking Rate (Words per Minute)
Histogram of 10 Feb 91 Test Speakers Speaking Rate
mean: 151.7
std_dev: 6.8
Number of Speakers
Speaking Rate (Words per Minute)
Histogram of 10 Sep 92 Test Speakers Speaking Rate
mean: 160.0
std_dev: 25.1
Number of Speakers
Speaking Rate (Words per Minute)

Figure 4

Yes. Some properties of this test set appear not to be representative of the properties of the training material (e.g., some dialect regions are over- or under-represented, and rate-of-speech properties indicate an unusual number of outliers). Ideally, all properties of all test sets should be representative of the training material.

Yes and No. Since all sites participating in this test used identical training and test material, valid comparisons across systems and sites may be made. However, those comparisons are specifically valid in the context of this test set, and may not generalize to other test sets.

No. Comparisons with earlier test sets may be invalid, since developers have had opportunities to "tune" their systems to the properties of earlier test sets, so that all previous test sets inevitably seem "easier" than they were when first used as test material. This is a well-known phenomenon in pattern recognition.

No. This may be a valid, but small sample, of a distribution of speakers with an inherently large variance. The test sets are in any event much too small (i.e., 10-12 speakers) to expect good sampling of a "speaker-independent" user population, and are also too small to show small performance increments with statistical significance. One must keep these considerations of statistical significance in mind when interpreting these results.

Lessons to be Learned

A number of lessons might be learned. Perhaps most importantly, one might well consider requiring larger speaker populations for tests of speaker-independent technology, so as to minimize effects due to statistical outliers. One might also propose using larger test sets in the later stages of research and technology development, so as to facilitate making statistically significant inferences about small incremental progress. And, one might pre-screen all potential test material with a state-of-the-art speech recognition system to identify any potential "outliers", and use this information to make the distribution of test material more uniform.

4. Statistical Significance Tests

When discussing the results of benchmark tests, it is wise to consider the statistical significance associated with the test results. In many cases, simple comparison of the differences between word error (or sentence error) rates, without use of statistical significance tests, may be misleading. Within the DARPA speech research community, use of two significance tests has become routine [8], following suggestion of these tests by Gillick and Cox [9]. Two new tests were implemented for the most recent tests, one of which was suggested for use by Makhoul at a DARPA ANN Technology Program meeting in the Spring of 1992.

The two tests used in previous tests are termed the MAtched-Pair Sentence Segment Word Error (MAPSSWE) test, and the McNemar sentence error rate test. The MAPSSWE test is a paired-comparison test which identifies corresponding sentence segments excised from sentence hypotheses for two speech recognition systems. Each segment is bounded by either the beginning or end of the sentence or by at least two correctly recognized words. The differences in word error rates for the corresponding segments are then compared. The McNemar test is another paired comparison test, acting on the sentence error data, directing attention to the size of the subsets of errors that are unique to each system.

The sign test, suggested for use in these tests by Makhoul, consists of a test on the individual word error rates for each test speaker. The proportion of the test speakers for which one system had a smaller word error rate than another system is identified by considering the sign of the difference between word error rates for each speaker, and comparing this proportion with 50%. The other new test used for these tests is the Wilcoxon signed-rank test, which is an extension of the sign test. Not only are the differences in word error rates considered, and the sign of the differences considered, but the differences are rank-ordered, and the sum of the signed ranks are determined. The probability of the observed sum of signed ranks can be estimated for a test of the hypothesis that the two populations (represented by the respective members of the matched pairs) are identical. Discussions of the sign test and the Wilcoxon signed-rank test can be found in the literature on nonparametric statistical tests (e.g., [10 - 14]).

For the current benchmark tests, NIST implemented four tests, designated "MN" for the McNemar sentence error rate test, "MP" for the Matched Pair sentence segment word error test, "SI" for the sign test, and "WI" for the WIlcoxon signed-rank test.

NIST's conventional practice has been to implement these tests for all possible cross-system comparisons, and to show the results in the form of a matrix with the word "same" printed for each test in which the null hypothesis is valid (i.e., the differences between systems are not demonstrated to be significant using the relevant test), and the identity of the system with the lower error rate in the event that the null hypothesis is shown to be invalid.

5. Summary of Test Results

Table 3 presents a tabulation of the results using the September 1992 test material for the word-pair grammar, only. Table 4 has the statistical significance test matrix for the speaker-independent tests using the word-pair grammar. Many other results were provided to test participants, and may be cited in other papers in this proceedings. Reference should be made to relevant papers for full descriptions of the technical approaches used for each system.

Table 3: Tabulated Benchmark Test Results for September 1992 Test Sets

1. System Descriptor Codes

att1	AT&T Lee-Chou-Juang SE System
att2	AT&T Lee-Gauvain SD System
att3	AT&T Ljolje-Riley SI System
bbn1	BBN BYBLOS HMM Baseline SI System
bbn2	BBN SNN/HMM Hybrid SI System
bu1	Boston U. BU SSM SI System
bu2	Boston U. BU SSM-BBN BYBLOS SI System
cam	Cambridge U. Recurrent Net SI System
cmu	Carnegie-Mellon U. Sphinx-II SI System
htk3	Cambridge U. HTK Var. Mixture Triphone SI System
htk2	Cambridge U. HTK 15 Mixture Monophone SI System
htk1	Cambridge U. HTK 6 Mixture Triphone SI System
limsi	LIMSI SI Recognizer
mit_lll	MIT Lincoln Lab SI System
mit_ll2	MIT Lincoln Lab SD System
philips1	Philips Non-tied Mixtures SD System
philips2	Philips Tied Mixtures SD System
philips3	Philips Non-Tied Mixtures SI System
sri1	SRI Pure HMM Baseline SI System
sri2	SRI Pure CI-MLP Baseline System
sri3	SRI Pure CD-MLP Hybrid System
sri4	SRI Mixed MLP/HMM Probabilities SI System

2. SPEAKER DEPENDENT SYSTEMS - WORD-PAIR GRAMMAR

	#UTT	% Corr	% Sub	% Del	% Ins	% W.E.	% U.E.
att2-wd	300	98.4	1.0	0.6	0.3	1.9	12.7
mit_ll2-wd	300	97.8	1.3	0.9	0.4	2.5	17.0
philips1-wd	300	97.7	1.6	0.6	0.3	2.6	16.3
philips2-wd	300	98.2	1.2	0.5	0.1	1.8	12.3

3. SPEAKER INDEPENDENT SYSTEMS - WORD-PAIR GRAMMAR

	#UTT	% Corr	% Sub	% Del	% Ins	% W.E.	% U.E.
att1-wi	300	95.2	3.4	1.4	0.5	5.4	29.3
att3-wi	300	92.5	5.6	1.9	2.6	10.1	38.0
bbn1-wi	300	94.9	3.8	1.3	1.6	6.7	30.3
bbn2-wi	300	95.3	3.6	1.0	1.4	6.1	29.3
bu1-wi	300	93.4	4.8	1.9	1.8	8.5	40.3
bu2-wi	300	94.4	3.8	1.8	1.3	7.0	33.3
cam-wi	300	90.0	7.7	2.3	1.8	11.7	42.3
cmu-wi	300	94.8	4.1	1.1	1.0	6.2	31.3
htk1-wi	300	93.6	4.4	2.0	1.0	7.4	36.0
htk2-wi	300	91.3	6.5	2.2	1.0	9.7	39.3
htk3-wi	300	90.3	6.5	3.2	1.8	11.4	48.0
limsi-wi	300	96.0	2.9	1.2	0.4	4.4	25.0
mit_lll-wi	300	93.9	4.1	2.0	1.3	7.5	32.0
philips3-wi	300	94.6	3.6	1.9	0.6	6.0	30.0
sri1-wi	300	91.4	6.5	2.1	1.5	10.1	40.0
sri2-wi	300	90.6	6.4	3.0	1.5	10.9	39.3
sri3-wi	300	93.9	4.5	1.6	1.5	7.7	30.7
sri4-wi	300	93.2	5.2	1.6	1.0	7.8	31.7

Composite Report of All Significance Tests
For the Sep 92 RM1 rm1_ind_wg_109 Test

Test Name	Abbrev.
Matched Pair Sentence Segment (Word Error) Test	MP
Signed Paired Comparison (Speaker Word Accuracy) Test	SI
Wilcoxon Signed Rank (Speaker Word Accuracy) Test	WI
McNemar (Sentence Error) Test	MN

	att1-wi	att3-wi	bbn1-wi	bbn2-wi	bol-wi	bs2-wi	cam-wi	cmu-wi	htk1-wi	htk2-wi	htk3-wi	limsi-wi	mit_lll-wi	philips2-wi	sri1-wi	sri2-wi	sri3-wi	sri4-wi
att1-wi		MP att1-wi SI same WI att1-wi MN att1-wi	MP same SI same WI att1-wi MN same	MP same SI same WI same MN same	MP att1-wi SI att1-wi WI att1-wi MN att1-wi	MP att1-wi SI same WI same MN same	MP att1-wi SI same WI att1-wi MN att1-wi	MP same SI same WI same MN same	MP att1-wi SI att1-wi WI att1-wi MN att1-wi	MP att1-wi SI att1-wi WI att1-wi MN att1-wi	MP att1-wi SI att1-wi WI att1-wi MN att1-wi	MP same SI same WI limsi-wi same	MP att1-wi SI same WI att1-wi same	MP same SI same WI same MN same	MP att1-wi SI att1-wi WI att1-wi MN att1-wi	MP att1-wi SI att1-wi WI att1-wi MN att1-wi	MP att1-wi SI same WI same MN same	MP att1-wi SI same WI same MN same
att3-wi			MP bbn1-wi SI same WI same MN bbn1-wi	MP bbn2-wi SI same WI bbn2-wi MN bbn2-wi	MP same SI same WI same MN same	MP bs2-wi SI same WI same MN same	MP same SI same WI same MN same	MP cmu-wi SI same WI same MN cmu-wi	MP htk1-wi SI same WI same MN same	MP same SI same WI same MN same	MP same SI same WI same MN att3-wi	MP limsi-wi SI same WI limsi-wi MN limsi-wi	MP mit_lll-wi SI same WI same MN mit_lll-wi	MP philips3-wi SI same WI philips3-wi MN philips3-wi	MP same SI same WI same MN same	MP same SI same WI same MN same	MP sri3-wi SI same WI same MN sri3-wi	MP sri4-wi SI same WI same MN sri4-wi
bbn1-wi				MP same SI same WI same MN same	MP bbn1-wi SI same WI bbn1-wi MN bbn1-wi	MP same SI same WI same MN same	MP same SI same WI same MN same	MP bbn1-wi SI bbn1-wi WI bbn1-wi MN bbn1-wi	MP same SI same WI same MN same	MP same SI same WI same MN same	MP bbn1-wi SI bbn1-wi WI bbn1-wi MN bbn1-wi	MP limsi-wi SI limsi-wi WI limsi-wi MN same	MP same SI same WI same MN same	MP same SI same WI same MN same	MP bbn1-wi SI bbn1-wi WI bbn1-wi MN bbn1-wi	MP same SI same WI same MN same	MP same SI same WI same MN same
bbn2-wi					MP bbn2-wi SI same WI bbn2-wi MN bbn2-wi	MP same SI same WI same MN same	MP bbn2-wi SI bbn2-wi WI bbn2-wi MN bbn2-wi	MP same SI same WI same MN same	MP same SI same WI same MN bbn2-wi	MP bbn2-wi SI bbn2-wi WI bbn2-wi MN bbn2-wi	MP same SI same WI same MN same	MP limsi-wi SI same WI limsi-wi MN same	MP same SI same WI same MN same	MP same SI same WI same MN same	MP bbn2-wi SI bbn2-wi WI bbn2-wi MN bbn2-wi	MP same SI same WI same MN same	MP bbn2-wi SI same WI same MN same
bol-wi						MP bs2-wi SI bs2-wi WI bs2-wi MN bs2-wi	MP bol-wi SI same WI same MN same	MP cmu-wi SI same WI cmu-wi MN cmu-wi	MP same SI same WI same MN same	MP same SI same WI same MN same	MP bol-wi SI same WI bol-wi MN bol-wi	MP limsi-wi SI limsi-wi WI limsi-wi MN limsi-wi	MP same SI same WI same MN mit_lll-wi	MP philips3-wi SI same WI philips3-wi MN philips3-wi	MP same SI same WI same MN same	MP bol-wi SI same WI same MN sri3-wi	MP same SI same WI same MN sri4-wi
bs2-wi							MP bs2-wi SI same WI bs2-wi MN bs2-wi	MP same SI same WI same MN same	MP same SI same WI same MN same	MP bs2-wi SI same WI bs2-wi MN same	MP bs2-wi SI bs2-wi WI bs2-wi MN bs2-wi	MP limsi-wi SI limsi-wi WI limsi-wi MN limsi-wi	MP same SI same WI same MN same	MP same SI same WI same MN same	MP bs2-wi SI bs2-wi WI same MN bs2-wi	MP same SI same WI same MN same	MP same SI same WI same MN same
cam-wi								MP cmu-wi SI cmu-wi WI cmu-wi MN cmu-wi	MP htk1-wi SI htk1-wi WI htk1-wi MN same	MP htk2-wi SI same WI htk2-wi MN same	MP same SI same WI same MN same	MP limsi-wi SI limsi-wi WI limsi-wi MN limsi-wi	MP mit_lll-wi SI mit_lll-wi WI mit_lll-wi MN mit_lll-wi	MP philips3-wi SI philips3-wi WI philips3-wi MN philips3-wi	MP same SI same WI same MN same	MP sri3-wi SI sri3-wi WI sri3-wi MN sri3-wi	MP sri4-wi SI sri4-wi WI sri4-wi MN sri4-wi
cmu-wi									MP same SI same WI cmu-wi MN same	MP cmu-wi SI cmu-wi WI cmu-wi MN cmu-wi	MP cmu-wi SI cmu-wi WI cmu-wi MN cmu-wi	MP limsi-wi SI same WI limsi-wi MN limsi-wi	MP same SI same WI same MN same	MP same SI same WI same MN same	MP cmu-wi SI cmu-wi WI cmu-wi MN cmu-wi	MP cmu-wi SI cmu-wi WI cmu-wi MN cmu-wi	MP same SI same WI same MN same	MP same SI same WI same MN same
htk1-wi									MP htk1-wi SI same WI same MN same	MP htk1-wi SI same WI htk1-wi MN htk1-wi	MP limsi-wi SI limsi-wi WI limsi-wi MN limsi-wi	MP same SI same WI same MN same	MP same SI same WI same MN same	MP same SI same WI same MN same	MP htk1-wi SI same WI htk1-wi MN same	MP htk1-wi SI same WI same MN same	MP same SI same WI same MN same	MP same SI same WI same MN same
htk2-wi										MP same SI same WI same MN htk2-wi	MP limsi-wi SI limsi-wi WI limsi-wi MN limsi-wi	MP same SI same WI same MN same	MP mit_lll-wi SI mit_lll-wi WI mit_lll-wi MN mit_lll-wi	MP philips3-wi SI philips3-wi WI philips3-wi MN philips3-wi	MP same SI same WI same MN same	MP sri3-wi SI same WI same MN sri3-wi	MP sri4-wi SI same WI same MN sri4-wi
htk3-wi											MP limsi-wi SI limsi-wi WI limsi-wi MN limsi-wi	MP same SI same WI same MN same	MP mit_lll-wi SI mit_lll-wi WI mit_lll-wi MN mit_lll-wi	MP philips3-wi SI philips3-wi WI philips3-wi MN philips3-wi	MP same SI same WI same MN sri1-wi	MP sri3-wi SI sri3-wi WI sri3-wi MN sri2-wi	MP sri4-wi SI sri4-wi WI sri4-wi MN sri4-wi
limsi-wi												MP limsi-wi SI same WI limsi-wi MN limsi-wi	MP limsi-wi SI same WI limsi-wi MN limsi-wi	MP limsi-wi SI same WI limsi-wi MN same	MP limsi-wi SI limsi-wi WI limsi-wi MN limsi-wi	MP limsi-wi SI limsi-wi WI limsi-wi MN limsi-wi	MP limsi-wi SI limsi-wi WI limsi-wi MN limsi-wi
mit_lll-wi													MP same SI same WI same MN same	MP mit_lll-wi SI mit_lll-wi WI mit_lll-wi MN mit_lll-wi	MP same SI same WI same MN same	MP same SI same WI same MN same
philips3-wi														MP philips3-wi SI philips3-wi WI philips3-wi MN philips3-wi	MP same SI same WI same MN same	MP philips3-wi SI same WI same MN same
sri1-wi															MP same SI same WI same MN same	MP sri3-wi SI same WI sri3-wi MN same	MP sri4-wi SI sri4-wi WI sri4-wi MN sri4-wi
sri2-wi																MP sri3-wi SI same WI sri3-wi MN sri3-wi	MP sri4-wi SI same WI sri4-wi MN sri4-wi
sri3-wi																	MP same SI same WI same MN same
sri4-wi																	

Table 4

Speaker Dependent Systems

Three sites participated in tests using the speaker dependent test material: AT&T Bell Laboratories, MIT Lincoln Laboratory, and Philips. Both the "No Grammar" and "Word Pair Grammar" were used for these tests. The word error for the case of the use of the "Word Pair Grammar" ranges from 2.6% to 1.8%.

The lowest error rate (1.8%) was reported for the system designated "philips2-wd", a tied mixture HMM system incorporating linear discriminant analysis on 3 consecutive 63 component vectors to yield a 35 dimension transformed vector as input to the system. The 991 word lexicon is partitioned into two subsets, one including about 100 short function words.

Implementation of the statistical significance tests for this system indicates that the differences in performance with another Philips system, philips1-nd, and with the att2-wd system are not shown to be significant with any of the 4 significance tests, although for all 4 tests, differences with the mit_ll2-wd system are shown to be significant.

Speaker Independent Systems

A total of 18 sets of results was reported for the interesting case of speaker independent systems using the September 1992 speaker independent test set. Nine sites were involved: AT&T Bell Laboratories, BBN, Boston University, Cambridge University, Carnegie Mellon University, CNRS-LIMSI, MIT Lincoln Laboratory, Philips, and SRI International. The word error for the case of the "Word Pair Grammar" ranges from 11.7% to 4.4%.

The lowest error rate (4.4%%) was reported for the system designated "limsi-wi", an HMM system incorporating a reduced set of 36 phones (including silence), for which each phone model is a left-to-right context dependent HMM, which associates pronunciation graphs with each word (so as to allow alternate pronunciations, including optional phones) and also incorporates word boundary phonologicial rules in building the phone graph used by the recognizer.

Implementation of the statistical significance tests for this system indicates that the differences in performance between the limsi-wi system and the bbn2-wi and philips3-wi systems are not shown to be significant with 2 of the 4 significance tests, the sign test and the McNemar sentence error rate test, although the other 2 tests indicate significant differences. All three of these systems perform very well. In many cases, the significance tests indicate significant differences in favor of the limsi-wi system when compared with other systems.

Hybrid HMM-ANN Systems

Two DARPA ANN Technology contractors participated in official benchmark tests using the RM1 corpus and test material: BBN and SRI International. For each site, comparisons

of results between "baseline" HMM systems and "hybrid" ANN-HMM systems were possible in order to make inferences about the incremental performance gains made possible by incorporation of ANN technology. Refer to other papers in this proceedings for descriptions of the approaches used by these contractors.

In the tabulation of results for speaker-independent systems with no-grammar condition, the baseline BBN system is designated bbn1, with 6.7% word error. The BBN hybrid system is designated bbn2, with 6.1% word error, about a 10% reduction in word error rate. None of the 4 significance tests, however, indicate this difference in performance between the baseline HMM and hybrid ANN-HMM system to be significant.

The SRI baseline HMM system is designated sri1, with 10.1% word error, which can be contrasted with the results for the hybrid system, sri4, with 7.8% word error. For the SRI systems, incorporation of ANN technology results in approximately a 20% reduction in word error. All 4 of the significance tests indicate that this difference is significant.

Note, however, that while the hybrid BBN system has lower word error than the hybrid SRI system, of the 4 significance tests implemented by NIST, only the matched-pair sentence-segment word error tests indicates that the difference is significant.

One conclusion supported by these analyses is that the test sets are, in general, too small, particularly with regard to the number of speakers, to reveal small differences in performance. This is a particularly problematic issue in using test sets of this size (i.e., 10 speakers) for developmental purposes, and some developers report having combined a number of previously released test sets into larger test sets for developmental test purposes. Other developers report having observed systematic trends toward reduced error rates while using small developmental test sets that, in many cases, result in significant improvements when tested with new test sets.

The fact that the speaker-independent test set speaker population for the September 1992 tests is somewhat unusual, with, as noted, a number of outliers with regard to rate of speech, is a further complicating factor.

6. Summary

This paper has documented the final Benchmark Test results for the DARPA Resource Management (RM1) continuous speech corpus, using the September 1992 Test Set(s).

Some properties of the speaker-independent test set are shown to be unusual (e.g., the distribution of speaking rate for the 10 speakers in the speaker-independent test set), and it has been hypothesized that this factor may account for somewhat higher reported error rates than for previous test sets. If this is so, it may be the case that current technology does not accommodate fast or slow rates of speech very well.

These tests included participation of European speech research groups, and the lowest reported error rates for both speaker-dependent and speaker-independent technologies

were reported by Philips Research Laboratories, in Aachen, Germany, and CNRS-LIMSI, in Paris, France, respectively.

The tests also included results for several hybrid ANN-HMM systems and provided comparisons with baseline HMM systems. Incorporation of ANN technology into HMM-based systems has been shown to yield improvements of between 10 to 20 percent reduction in word error.

Performance differences, in many cases, between well-performing systems using this test material are not shown to be significant. There are a number of well-performing systems, and comparisons need to be considered on a case-by-case basis using a number of different significance tests.

Acknowledgement

At NIST, Alvin Martin and Bill Fisher developed hypotheses to address questions regarding the degree of difficulty of the September 1992 test set. We acknowledge with gratitude the cooperation of many individuals at the participating site in locally implementing the tests, in preparing systems descriptions for distribution at the meeting, and in sending the results to NIST in a timely manner, and the generous support and patience of Dr. Barbara Yoon, DARPA ANN Technology Program Manager.

Disclaimer

The data summarized in this paper was derived at NIST by uniform implementation of scoring software, operating on benchmark test results provided by individual contractors and cooperating research organizations for unsupervised, locally implemented tests. The results of tests conducted with this test material and analyses of performance are not to be construed as official findings of NIST, the Department of Commerce, DARPA, the Department of Defense, or the United States Government. No endorsement of any systems or algorithms are intended. In many cases the statistical significance of differences in the reported results for different systems can not be determined, partially because of limitations in the size and composition of the test sets.

References

[1] Fisher, W.M., "The DARPA Task Domain Speech Recognition Database", Speech Recognition: Proceedings of a Workshop, San Diego, CA, March 19-20, 1987, Science Applications International Corporation Report SAIC-87/1644, pp. 105-109.

[2] Price, P., et al., "The DARPA 1000-Word Resource Management Database for Continuous Speech Recognition", Paper S.13.21 in Proceedings of the 1988 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 651-654.

[3] Pallett, D.S., "Test Procedures for the March 1987 DARPA Benchmark Tests", Speech Recognition: Proceedings of a Workshop, San Diego, CA, March 19-20, 1987, Science Applications International Corporation Report SAIC-87/1644, pp. 75-78.

[4] Pallett, D.S., "Selected Test Material for the March 1987 DARPA Benchmark Tests", Speech Recognition: Proceedings of a Workshop, San Diego, CA, March 19-20, 1987, Science Applications International Corporation Report SAIC-87/1644, pp. 79-81.

[5] Pallett, D.S. "Speech Results on Resource Management Task", Proceedings of Speech and Natural Language Workshop", Philadelphia, PA, February 21-23, 1989, ISBN 1-55860-073-6, Morgan Kaufmann Publishers, Inc., pp. 18-24.

[6] Pallett, D.S., "DARPA Resource Management Benchmark Test Results: June 1990", Proceedings of Speech and Natural Language Workshop", Hidden Valley, PA, June 24-27, 1990, ISBN 1-55860-157-0, Morgan Kaufmann Publishers, Inc., pp. 298-305.

[7] Pallett, D.S., "Session 2: DARPA Resource Management and ATIS Benchmark Test Poster Session", Proceedings of Speech and Natural Language Workshop", Pacific Grove, CA, February 19-22, 1991, ISBN 1-55860-207-0, Morgan Kaufmann Publishers, Inc., pp. 49-58.

[8] Pallett, D.S., "Tools for the Analysis of Benchmark Speech Recognition Tests", Paper S.2.16 in Proceedings of the 1990 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 97-100.

[9] Gillick, L. and Cox, S., "Some Statistical Issues in the Comparison of Speech Recognition Algorithms", Paper S. 10.b.5 in Proceedings of the 1989 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 532-535.

[10] Winkler, R.L. and Hayes, W.L., "Statistics: Probability, Inference, and Decision", Second Edition, Holt, Rinehart and Winston, New York, NY, 1975.

[11] Lehmann, E.L., "Nonparametrics: Statistical Methods Based on Ranks", Holden-Day Inc., San Francisco, CA, 1975.

[12] Gibbons, J.D., "Nonparametric Statistical Inference", Second Edition, Marcel Dekker, Inc., New York, 1985.

[13] Book, S. A., "Statistics: Basic Techniques for Solving Applied Problems", McGraw-Hill, New York, 1977.

[14] Bradley, J.V., "Distribution Free Statistical Tests", Prentice-Hall, Inc., Englewood Cliffs, NJ, 1967.